University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2008

Tools for detection of protein interactions in biomedical literature

Rinaldi, Fabio ; Schneider, G ; Kaljurand, K

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-8813
Conference or Workshop Item

Originally published at:
Rinaldi, Fabio; Schneider, G; Kaljurand, K (2008). Tools for detection of protein interactions in biomedical literature. In: Genomes to Systems Conference 2008, Manchester, UK, 17 March 2008 - 19 March 2008.

Tools for Detection of Protein Interactions in Biomedical Literature

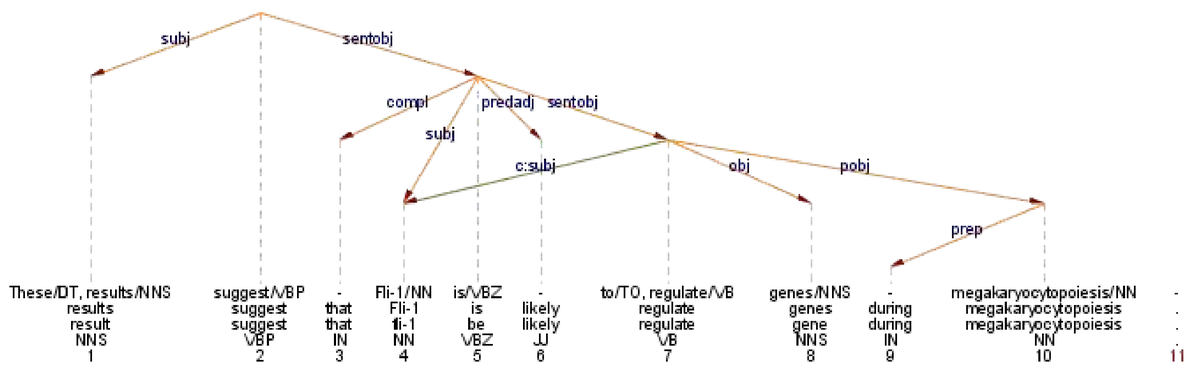
Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand
Institute of Computational Linguistics
University of Zurich
www.ontogene.org

Introduction

The biomedical literature is a treasure trove of results which can potentially support the R&D process, by suggesting new research targets, or by preventing the duplication of already performed experiments. However, often such potential is left untapped due to the lack of tools which can be effectively adopted in the process of literature-based discovery. We present techniques which can support this process, focusing in particular on the detection of interactions between biomedical entities (genes, diseases, proteins, etc.).

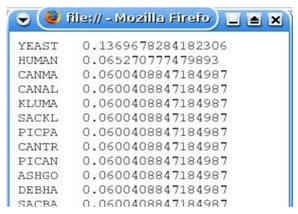
Methods

Our approach is based on a dependency parser and modular rules which make use of rich linguistic annotations. Our system is organized around a pipeline of NLP tools, which perform tasks such as sentence splitting, tokenization, PoS tagging, lemmatization, term extraction, chunking, dependency parsing. The final result of this stage of analysis is a set of dependencies.

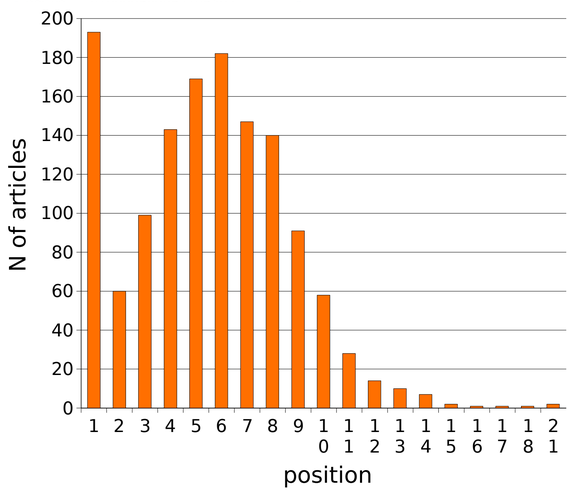


The task of detecting protein-protein interactions is additionally complicated by the need of locating protein names in the articles and normalizing them to unique identifiers (e.g. from UniProt). Further, not all potential interactions are requested, but only those that the authors present as their main results. The combination of these requirements makes the problem extremely challenging.

The Cap - binding protein eIF4E promotes folding of a functional domain of yeast translation initiation factor eIF4G1 .
The association of eucaryotic tra IF4E_YEAST tion factor eIF4G with the cap - binding protein eIF4E establishes a critical link between the mRNA and the ribosome during translation initiation .
This association requires a conserved seven amino acid peptide within eIF4G that binds to eIF4E .
Here we report that a 98 - amino acid fragment of S . cerevisiae eIF4G1 that contains this eIF4E binding peptide undergoes an unfolded to folded transition upon binding to eIF4E .
The folding of the eIF4G1 domain was evidenced by the eIF4E - dependent changes in its protease sensitivity and (1) H - (15) N HSQC NMR spectrum .
Analysis of a series of charge - to - alanine mutations throughout the essential 55.4 - kDa core of yeast eIF4G1 also revealed substitutions within this 98 - amino acid region that led to reduced eIF4E binding in vivo and in vitro .
These data suggest that the association of yeast eIF4E with eIF4G1 leads to the formation of a structured domain within eIF4G1 that could serve as a specific site for interactions with other components of the translational apparatus .
They also suggest that the stability of the native eIF4E - eIF4G complex is determined by amino acid residues outside of the conserved seven - residue consensus sequence .



Ambiguity in protein names is a well-known and widespread problem. Being able to determine with precision which is the organism used in the study leads to a significant disambiguation effect.



Since not all the interactions reported by the authors are relevant for the curation process, it is necessary to identify reliably sentences which present the authors' own results. We have adopted an efficient 'novelty' filter, which can distinguish background from novel knowledge.

Relation mining is based on cascading rules, which are organized modularly in order to support increasingly abstract types of queries.

sid	Sentence
m92013023-s1 SVG	Anti-CD2 receptor antibodies activate the HIV long terminal repeat in T lymphocytes .
m91355651-s5 SVG	We found that in both cell lines , both phorbol ester and TNF alpha were able to activate NF-kappa B .
m91355651-s5 SVG	We found that in both cell lines , both phorbol ester and TNF alpha were able to activate NF-kappa B .
m94148994-s9 SVG	These data suggest that interferon regulatory factor 1 not only triggers the activation of the interferon signal transduction pathway , but also may play a role in limiting the duration of this response by activating the transcription of IRF-2 .
m92107162-s5 SVG	The simian virus 40 early promoter is also synergistically activated by the Z/c-myb combination .
m91237803-s2 SVG	Human herpesvirus 6 (HHV-6) can activate the human immunodeficiency virus (HIV) promoter and accelerate cytopathic effects in HIV-infected human T cells .

Evaluation

The results have been validated on a publicly available corpus [1] and by participation to a text mining competition (BioCreative) [2].

In the case of BioCreative, our system, after generating candidate interactions on the basis of co-occurrence of protein names within the same sentence, applies a novelty filter and a syntactic filter in order to separate meaningful interactions from accidental ones. The results, which are among the best reported, prove the effectiveness of the approach.

BioCreative Protein Interaction task

SUBMISSION		ARTICLES EVALUATED			ARTICLES WITH PREDICTIONS			OVERALL INTERACTOR		
TEAM	RUN	P	R	F-score	P	R	F-score	P	R	F-score
36	1	0.1422	0.2731	0.1740	0.1753	0.3365	0.2144	0.1039	0.2467	0.1462
36	2	0.0974	0.2875	0.1329	0.1132	0.3340	0.1544	0.0655	0.2723	0.1057
36	3	0.1527	0.2415	0.1705	0.2003	0.3167	0.2236	0.1175	0.2165	0.1523
40	1	0.2355	0.4110	0.2789	0.2724	0.4753	0.3225	0.1544	0.3549	0.2152
40	2	**0.3720**	0.3491	**0.3386**	0.4960	0.4654	0.4514	0.3246	0.3047	**0.3143**
42	1	0.0699	**0.6313**	0.1200	0.0769	**0.6939**	0.1319	0.0402	**0.6161**	0.0755
42	2	0.2809	0.2631	0.2559	0.4451	0.4169	**0.4055**	0.2792	0.2455	0.2613
42	3	0.2748	0.2693	0.2560	0.4190	0.4108	0.3904	0.2677	0.2489	0.2580

Conclusion

Although fully automated extraction of interactions is still not within immediate reach, recent results show that our tools already perform at a competitive level, making them interesting either as stand-alone modules for preliminary document inspection, or as components within an environment aimed at supporting the process of curation of biomedical literature.

Acknowledgments

This research is partially supported by the Swiss National Science Foundation (grant 100014-118396/1).

References

[1] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, and Martin Romacker, *An Environment for Relation Mining over Richly Annotated Corpora: the case of GENIA*, BMC Bioinformatics, 7(Suppl 3), S3, (2006).

[2] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon, *Ontogene in Biocreative II*, Genome Biology, (2008). (to appear).